

MATHEMATICS AND IMPLEMENTATION DETAILS OF A BLOCK MINRES ALGORITHM*

KIRK M. SOODHALTER[†]

Abstract. We present a block minimum residual (MINRES) algorithm for symmetric indefinite matrices based on an alternate method for constructing the block Lanczos vectors. This method allows us to compare performance with the standard MINRES algorithm iteration for iteration, and it handles removal of dependent Lanczos vectors more gracefully. We describe both a theoretical derivation of the algorithm as well as practical implementation details. Some numerical results are shown to illustrate performance on some sample problems. We also present some experiments to show how the relationship between right-hand sides affects the performance of this method.

1. Introduction. We wish to efficiently solve

$$\mathbf{A}\mathbf{X} = \mathbf{B} \tag{1.1}$$

were $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a symmetric, indefinite matrix, and $\mathbf{B} \in \mathbb{R}^{n \times p}$. If $p = 1$, preconditioned Krylov subspace iterative methods for symmetric systems [11, 14] have been shown to be effective for solving such problem. For the case $p > 1$, there are several options. We can solve for each right-hand side in sequence, ignoring the relationship between the systems. In solving this sequence of systems, we can also take advantage of the fact that the matrix does not change and employ deflation techniques e.g., [1, 16, 20, 24], or other projection techniques, e.g., [5, 18, 22], to accelerate convergence of an iteration on one system by deflating with a subspace generated during the previous system's iteration. Moreover, we can solve these systems in parallel, by solving each system on a different processor.

We can also employ block Krylov subspace methods. For solving a linear system with multiple right-hands, it is quite natural to describe iterative methods which extend the idea of the Krylov subspace to the block setting, in which we have more than one starting vector. In the case of symmetric systems, so-called block extensions of methods such as conjugate gradients [11] and the minimum residual method (MINRES) [14], have been previously described [13]. For the case $p = 1$, block methods can also be used as a way to accelerate the convergence of the iteration, where we generate random right-hand sides with which to generate a basis for the block Krylov subspace over which we minimize. Thus, block methods have utility for all $p \in \mathbb{N}$.

In this paper, we present an implementation of the block MINRES algorithm, different than the one presented by O'Leary in, [13]. This implementation is based on an alternative method of generating a block Krylov subspace introduced by Ruhe [17] and further described (for the non symmetric case) in [19, Section 6.12]. This algorithm can be considered a simplification of the algorithm presented in [2] in the case that A is symmetric, which extends Ruhe's method to generalize the nonsymmetric Lanczos process to the block setting. We present both the theoretical details and the practical implementation issues for this algorithm. To our knowledge, this is the first paper to provide the implementation details of a block minimum residual algorithm for symmetric matrices *.

[†]Industrial Mathematics Institute, Johannes Kepler University, Altenbergerstraße 69, A-4040 Linz, Austria. (kirk.soodhalter@indmath.uni-linz.ac.at)

*This version dated January 11, 2013.

*Matlab implementation available at <http://math.soodhalter.com/software.php>

In the next section, we introduce notation and review relevant theory about block Krylov subspace methods. In Section 3, we derive a version of the block minimum residual method built upon Ruhe’s block Lanczos method. In Section 4, we derive our new version of the block MINRES algorithm in detail and describe our implementation, built to take advantage of the potential memory savings afforded by the method. Special attention is given to how data is stored to keep the scheme as simple as possible. We also present modifications to our implementation which accommodate the occurrence of dependence of the block Krylov subspace basis. In Section 5, we briefly discuss convergence properties of block methods. In Section 6, we present numerical results.

2. Preliminaries. We begin by describing some nomenclature and notation. We call a vector with multiple columns, such as \mathbf{B} when $p > 1$, a *block vector*. If we want to indicate the number p of columns, we will identify the vector as being block- p . Boldface, upper-case letters will be used to denote matrices, including block vectors. Boldface lower-case letters will denote column vectors. We denote the Euclidean norm by $\|\cdot\|$ and the Frobenius norm by $\|\cdot\|_F$. For a square, nonsingular matrix \mathbf{A} , we will denote the condition number associated to the 2-norm, $\kappa(\mathbf{A}) = \|\mathbf{A}\| \|\mathbf{A}^{-1}\|$. When identifying an equation as a QR-factorization, we will use the convention that the right-hand side of the equation is the QR-factorization of the left-hand side of the equation. We denote the $k \times k$ identity matrix \mathbf{I}_k . We also use the Matlab-inspired notation $(\cdot)_{i:j}$ to indicate rows i to j of the argument. For a matrix \mathbf{M} , we denote its range by $\mathcal{R}(\mathbf{M})$.

We also take a moment to clarify the ambiguity surrounding the word **deflation**. This term refers both to a class of techniques in which we project a Krylov subspace orthogonal to a specially selected subspace and construct approximations over an augmented Krylov subspace. This term also refers to the process of eliminating a dependent basis vector from a block Krylov subspace. In this work, when we use the term *deflation*, we mean the process used in augmented Krylov subspace methods. We will refer to the latter process simply as *removal of dependent basis vectors*.

Much has been written about the solution of linear systems with multiple right-hand sides. Various strategies have been presented. We could simply solve the systems in sequence, ignoring their relationship. Building upon this strategy, there has been extensive discussion about solving in sequence, but using the Krylov subspace information from one system to accelerate the solution of the next, either by deflated or recycled Krylov subspace methods [15, 18, 20]. Block Krylov subspace methods have been proposed in the symmetric case [12, 13, ?] and the nonsymmetric case, e.g., [9, 18]. In many cases, extending the framework of a Krylov method to the block right-hand side setting involves generalizing the machinery of, e.g., the Arnoldi process to deal with block vectors, see, for example, [19, Page 208]. We reproduce the algorithm here for convenience as Algorithm 2.1, which, at step j , generates a block Krylov subspace

$$\mathbb{K}_j(\mathbf{A}, \mathbf{V}_1) = \mathcal{K}_j(\mathbf{A}, \mathbf{v}_1^{(1)}) + \mathcal{K}_j(\mathbf{A}, \mathbf{v}_1^{(2)}) + \cdots + \mathcal{K}_j(\mathbf{A}, \mathbf{v}_1^{(p)})$$

spanned by the columns of

$$\mathbf{W}_j = [\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_j],$$

where $\mathbf{V}_\ell = [\mathbf{v}_\ell^{(1)}, \mathbf{v}_\ell^{(2)}, \dots, \mathbf{v}_\ell^{(p)}]$ for $\ell = 1, \dots, j$.

Algorithm 2.1: Block Arnoldi Method

Input : $\mathbf{A} \in \mathbb{C}^{n \times n}$, $\mathbf{V}_1 \in \mathbb{C}^{n \times p}$, $\mathbf{V}_1^* \mathbf{V}_1 = \mathbf{I}_p$
Output: An orthonormal basis for $\mathcal{K}_m(\mathbf{A}, \mathbf{V}_1)$ and $\bar{\mathbf{H}}_j \in \mathbb{C}^{(j+1)p \times jp}$, $\bar{\mathbf{H}}_j$ has p lower subdiagonal entries

```

1 for  $j = 1, 2, \dots, m$  do
2   Compute  $\mathbf{W} = \mathbf{A}\mathbf{V}_j$ 
3   for  $i = 1, 2, \dots, j$  do
4      $\mathbf{H}_{i,j} \leftarrow \mathbf{V}_i^T \mathbf{W}$ 
5      $\mathbf{W} \leftarrow \mathbf{W} - \mathbf{V}_i \mathbf{H}_{i,j}$ 
6   Compute the QR-factorization  $\mathbf{W} = \mathbf{V}_{j+1} \mathbf{H}_{j+1,j}$ 

```

Normalization in the case of a single vector is generalized to the computation of an orthonormal basis for the column space of a block vector. For clarity, we will refer to methods that perform operations at the level of a block of vectors as *block-level Krylov methods*. Let $\mathbf{H}_j = (\mathbf{H}_{i,\ell})$ be generated by the block Arnoldi process and $\mathbf{H}_{i,\ell} \in \mathbb{C}^{p \times p}$. As a consequence of the block normalization, $\mathbf{V}_\ell \in \mathbb{C}^{n \times p}$ has orthonormal columns and the columns in each block are orthonormal to the columns of the other blocks; in other words, $\mathbf{W}_j^T \mathbf{W}_j = \mathbf{I}_j$. Let \mathbf{E}_j be the matrix that contains the last p columns of the \mathbf{I}_{jp} . As in the single vector case, we have a block Arnoldi relation

$$\mathbf{A}\mathbf{W}_j = \mathbf{W}_j \mathbf{H}_j + \mathbf{V}_{j+1} \mathbf{H}_{j+1,j} \mathbf{E}_j^*.$$

To derive a different MINRES algorithm, we will use the block Arnoldi algorithm proposed by Ruhe [17] and described in [19, Page 209]. For this method, we adopt the notation that $\mathbf{U}_p = [\mathbf{u}_1, \dots, \mathbf{u}_p]$ denotes the normalized starting block of vectors. Note that \mathbf{U}_p was called \mathbf{V}_1 when describing the block-level method. Essentially, Ruhe's method performs the same orthogonalization as the true block method, only one vector at a time. Therefore, at each iteration of Ruhe's block Arnoldi method, one matrix-single-vector product is performed as opposed to a matrix-block-vector product in a block-level method. For review, we present Ruhe's block Arnoldi process as Algorithm 2.2, as described in [19, Page 209].

Algorithm 2.2: Ruhe's Block Arnoldi Method

Input : $\mathbf{A} \in \mathbb{C}^{n \times n}$, $\mathbf{U}_p \in \mathbb{C}^{n \times p}$, $\mathbf{U}_p^* \mathbf{U}_p = \mathbf{I}_p$
Output: $\mathbf{U}_{p+j} \in \mathbb{C}^{n \times (m+p)}$, $\mathbf{U}_{p+j}^* \mathbf{U}_{p+j} = \mathbf{I}_{p+j}$ and $\bar{\mathbf{H}}_j \in \mathbb{C}^{(j+p) \times j}$, $\bar{\mathbf{H}}_j$ has p lower subdiagonal entries

```

1 for  $j = p, p+1, \dots, j+p-1$  do
2   Set  $k := j - p + 1$ 
3   Compute  $\mathbf{w} := \mathbf{A}\mathbf{u}_k$  for  $i = 1 : j$  do
4      $h_{i,k} := \mathbf{u}_i^* \mathbf{w}$ 
5      $\mathbf{w} \leftarrow \mathbf{w} - h_{i,k} \mathbf{u}_i$ 
6   Compute  $h_{j+1,k} := \|\mathbf{w}\|_2$  and  $\mathbf{u}_{j+1} := \mathbf{w}/h_{j+1,k}$ 

```

We can observe that at iteration j , where $j = \ell p$ is a multiple of the block size, Algorithm 2.2 has produced the same orthonormal basis as Algorithm 2.1 at step ℓ .

We also have Ruhe's block Arnoldi relation

$$\mathbf{A}\mathbf{U}_j = \mathbf{U}_{j+p}\overline{\mathbf{H}}_j. \quad (2.1)$$

The block Hessenberg matrix $\overline{\mathbf{H}}_j$ has p lower subdiagonal bands and has the structure

$$\overline{\mathbf{H}}_j = \begin{bmatrix} \mathbf{H}_j \\ \mathbf{H}_{p \times j} \end{bmatrix}$$

where \mathbf{H}_j is a square $j \times j$ matrix which is banded lower subdiagonal with p bands satisfying the identity

$$\mathbf{H}_j = \mathbf{U}_j^* \mathbf{A} \mathbf{U}_j. \quad (2.2)$$

Observe that $\mathbf{H}_{p \times j}$ only has nonzero entries in the last p columns with structure

$$\mathbf{H}_{p \times j} = \begin{bmatrix} \mathbf{0}_{p \times (m-p)} & \mathbf{C}_j \end{bmatrix}$$

where $\mathbf{C}_j \in \mathbb{C}^{p \times p}$ is upper triangular. At step j , we will use the following notation to describe the space we have constructed. We can always write $j = kp + m$ where $0 < j < p$. For the triple (j, k, m) , j indicates the current iteration, k indicates the beginning dimension of the constituent Krylov subspaces, and m indicates which subspace was increased from dimension k to dimension $k + 1$ at iteration j . To illustrate, the subspace that has been generated at step j is

$$\mathbb{K}_{j,k,m}(\mathbf{A}, \mathbf{U}_p) = \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{u}_1) + \cdots + \underbrace{\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{u}_m)}_{+1 \text{ dim. at iter. } j} + \underbrace{\mathcal{K}_k(\mathbf{A}, \mathbf{u}_{m+1})}_{+1 \text{ dim. at next iter.}} + \cdots + \mathcal{K}_k(\mathbf{A}, \mathbf{u}_p) \quad (2.3)$$

Similar to the Hermitian Lanczos relation in the case of a single-vector Krylov method, observe that if \mathbf{A} is Hermitian, the relation (2.2) implies that \mathbf{H}_j is also Hermitian. Due to the banded lower subdiagonal structure of \mathbf{H}_j , we see that \mathbf{H}_j is $2p + 1$ banded matrix with p superdiagonal entries and p subdiagonal entries. This structure implies that we only need the previous $2p$ basis vectors of $\mathbb{K}_{j,k,m}(\mathbf{A}, \mathbf{U}_p)$ in order to compute \mathbf{u}_{j+1} . We have the $2p + 1$ term recurrence relation

$$\mathbf{h}_{j+p,j} \mathbf{u}_{j+1} = \mathbf{A} \mathbf{u}_j - \sum_{\ell=\min(1, m-p)}^{m+p-1} h_{j,\ell} \mathbf{u}_\ell \quad (2.4)$$

Due to symmetry, we do not need to compute $h_{i,j}$ where $i < j$ since it was computed previously as $h_{j,i}$. This will require the storage of the lower subdiagonal entries of the last p columns of $\overline{\mathbf{H}}_j$. This yields Ruhe's block Lanczos method, which we present as Algorithm 2.2.

3. A Block Minimum Residual Method. We derive a minimal residual algorithm based on this version of the Lanczos algorithm. We take a few moments to describe what is meant by a block minimum residual algorithm. If we begin with an initial guess $\mathbf{X}_0 = \mathbf{0}$, at the j th step the following method will produce an approximation $\mathbf{X}_j \in \mathbb{C}^{n \times p}$ such that for each $0 < i \leq p$, the residual $\|\mathbf{b}^{(i)} - \mathbf{A} \mathbf{x}_j^{(i)}\|$ is minimized over the space $\mathbb{K}_{j,k,m}(\mathbf{A}, \mathbf{B})$, where $\mathbf{x}_j^{(i)}$ is the i th column of \mathbf{X}_j .

At step j , we want to minimize each column of the block residual $\mathbf{F}_j = \mathbf{B} - \mathbf{A}\mathbf{X}_j$ over $\mathbb{K}_{j,k,m}(\mathbf{A}, \mathbf{B})$. Following the explanation of MINRES presented in [8], we can derive a block MINRES algorithm based on Ruhe's block Lanczos process. Let $\mathbf{E}_1^{(j)} \in \mathbb{R}^{(j+p) \times p}$ be the matrix containing the first p columns of $\mathbf{I}_{(j+p)}$. Observe that

$$\mathbf{E}_1^{(j)} = \begin{bmatrix} \mathbf{E}_1^{(j-1)} \\ \mathbf{0}_{1 \times p} \end{bmatrix}. \quad (3.1)$$

Given an initial residual \mathbf{F}_0 we can normalize it using the QR factorization $\mathbf{F}_0 = \mathbf{U}_p \mathbf{S}$. At step j of Ruhe's block Lanczos process, we have the QR factorization $\overline{\mathbf{H}}_j = \mathbf{Q}_j \overline{\mathbf{R}}_j$ is such that $\mathbf{Q}_j \in \mathbb{C}^{(j+p) \times (j+p)}$ is unitary, and $\overline{\mathbf{R}}_j \in \mathbb{C}^{(j+p) \times j}$ is upper triangular. The matrix $\overline{\mathbf{R}}_j$ has a simple block structure,

$$\overline{\mathbf{R}}_j = \begin{bmatrix} \mathbf{R}_j \\ \mathbf{0}_{p \times j} \end{bmatrix},$$

where \mathbf{R}_j is a square, upper triangular, $j \times j$ matrix. Let $\mathbf{f}_j^{(i)}$ be the i th column of \mathbf{F}_j , the j th block residual. The minimization of $\|\mathbf{f}_j^{(i)}\|$ can be rewritten as

$$\begin{aligned} \|\mathbf{f}_j^{(i)}\| &= \min_{\mathbf{x} \in \mathbf{x}_0^{(i)} + \mathbb{K}_{j,k,m}(\mathbf{A}, \mathbf{B})} \|\mathbf{b}^{(i)} - \mathbf{A}\mathbf{x}\| \\ &= \min_{\mathbf{y} \in \mathbb{R}^j} \|\mathbf{f}_0^{(i)} - \mathbf{A}\mathbf{U}_j \mathbf{y}\| \\ &= \min_{\mathbf{y} \in \mathbb{R}^j} \|\mathbf{U}_{j+p} \mathbf{f}_0^{(i)} - \mathbf{U}_{j+p} \mathbf{A} \mathbf{U}_j \mathbf{y}\| \\ &= \min_{\mathbf{y} \in \mathbb{R}^j} \|\mathbf{U}_{j+p} \mathbf{f}_0^{(i)} - \mathbf{U}_{j+p} \overline{\mathbf{H}}_j \mathbf{y}\| \\ &= \min_{\mathbf{y} \in \mathbb{R}^j} \|\mathbf{f}_0^{(i)} - \overline{\mathbf{H}}_j \mathbf{y}\| \\ &= \min_{\mathbf{y} \in \mathbb{R}^j} \|\mathbf{Q}_j^* \mathbf{f}_0^{(i)} - \overline{\mathbf{R}}_j \mathbf{y}\|. \end{aligned} \quad (3.2)$$

The solution of this least squares problem can be computed by solving the normal equations,

$$\begin{aligned} \|\mathbf{f}_0^{(i)}\| \overline{\mathbf{R}}_j^* \mathbf{Q}_j^* e_{j+p}^{(i)} &= \mathbf{R}_j^* \mathbf{R}_j \mathbf{y}_j^{(i)} \\ \|\mathbf{f}_0^{(i)}\| \mathbf{R}_j^* \overline{\mathbf{R}}_j^* \mathbf{Q}_j^* e_{j+p}^{(i)} &= \mathbf{R}_j \mathbf{y}_j^{(i)} \\ \|\mathbf{f}_0^{(i)}\| [\mathbf{I}_j \quad \mathbf{0}_{k \times p}] \mathbf{Q}_j^* e_{j+p}^{(i)} &= \mathbf{R}_j \mathbf{y}_j^{(i)} \\ \|\mathbf{f}_0^{(i)}\| (\mathbf{Q}_j^* e_{j+p}^{(i)})_{1:j} &= \mathbf{R}_j \mathbf{y}_j^{(i)}. \end{aligned} \quad (3.3)$$

We can solve the normal equations individually for each right-hand side, or we can solve for all right-hand sides simultaneously, i.e., $\mathbf{Y}_j = \mathbf{R}_j^{-1}(\mathbf{Q}_j^* \mathbf{E}_1^{(j)} \mathbf{S})_{1:j}$, where \mathbf{S} replaces the individual initial residual norm $\|\mathbf{f}_0^{(i)}\|$ in the block setting. The block minimum residual approximation at step j is

$$\begin{aligned} \mathbf{X}_j &= \mathbf{X}_0 + \mathbf{U}_j \mathbf{Y}_j \\ &= \mathbf{X}_0 + \mathbf{U}_j \mathbf{R}_j^{-1}(\mathbf{Q}_j^* \mathbf{E}_1^{(j)} \mathbf{S})_{1:j}. \end{aligned} \quad (3.4)$$

Let $\bar{\mathbf{T}}_j = \mathbf{Q}_j^* \mathbf{E}_1^{(j)} \mathbf{S}$. We define our search directions as the columns of $\mathbf{M}_j = \mathbf{U}_j \mathbf{R}_j^{-1}$ and observe that the columns of \mathbf{M}_j successively span the same subspaces as the columns of \mathbf{U}_j due to the upper triangular structure of \mathbf{R}_j^{-1} . We denote block vector of search direction coordinates $\mathbf{T}_j = (\bar{\mathbf{T}}_j)_{1:j}$.

4. Block MINRES for Symmetric Linear Systems. To obtain a storage-efficient block MINRES algorithm based on Ruhe's block Lanczos method, we must discuss the structure of \mathbf{R}_j . This matrix is the upper $j \times j$ block of $\bar{\mathbf{R}}_j$ which is obtained from the QR-factorization of $\bar{\mathbf{H}}_j$. We can obtain this factorization column-by-column using Givens rotations to annihilate the subdiagonal entries of each column. Recall that for the column pair

$$\mathbf{h} = \begin{bmatrix} h_{i,j} \\ h_{i+1,j} \end{bmatrix},$$

to annihilate $h_{i+1,j}$, we can construct the Givens sine/cosine pair

$$s_{i+1,j} = \frac{h_{i+1,j}}{\sqrt{h_{i,j}^2 + h_{i+1,j}^2}} \quad \text{and} \quad c_{i+1,j} = \frac{h_{i,j}}{\sqrt{h_{i,j}^2 + h_{i+1,j}^2}}.$$

Annihilating $h_{i+1,j}$ can now be accomplished by premultiplication with a unitary matrix,

$$\begin{bmatrix} c_{i+1,j} & s_{i+1,j} \\ s_{i+1,j} & -c_{i+1,j} \end{bmatrix} \mathbf{h} = \begin{bmatrix} * \\ 0 \end{bmatrix}.$$

In column j , the Givens rotations annihilating the lower subdiagonal entries will affect rows j to $j+p$. Since $\bar{\mathbf{H}}_j$ has p superdiagonal entries, the Givens rotations associated to column j will add no more than p additional subdiagonal entries to any other column. Thus \mathbf{R}_j is upper triangular with $(2p+1)$ -bands. The j th column of \mathbf{R}_j has the following structure, denoted $\mathbf{r}_j^{(j)}$,

$$\mathbf{r}_j^{(j)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ r_{j-2p,j} \\ r_{j-2p+1,j} \\ \vdots \\ r_{j,j} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

We have that $\mathbf{M}_j = \mathbf{V}_j \mathbf{R}_j^{-1}$ so that $\mathbf{M}_j \mathbf{R}_j = \mathbf{V}_j$, and this gives us the relationship between the block Lanczos vectors and the search directions,

$$\begin{aligned}
r_{1,1} \mathbf{m}_1 &= \mathbf{v}_1 \\
r_{1,2} \mathbf{m}_1 + r_{2,2} \mathbf{m}_2 &= \mathbf{v}_2 \\
&\vdots \\
r_{1,2p+1} \mathbf{m}_1 + r_{2,2p+1} \mathbf{m}_2 + \cdots + r_{2p+1,2p+1} \mathbf{m}_{2p+1} &= \mathbf{v}_{2p+1} \\
r_{2,2p+2} \mathbf{m}_2 + r_{3,2p+2} \mathbf{m}_3 + \cdots + r_{2p+2,2p+2} \mathbf{m}_{2p+2} &= \mathbf{v}_{2p+2} \\
&\vdots \\
r_{j-2p,j} \mathbf{m}_{j-2p} + r_{j-2p+1,j} \mathbf{m}_{j-2p+1} + \cdots + r_{j,j} \mathbf{m}_j &= \mathbf{v}_j.
\end{aligned} \tag{4.1}$$

Thus, to compute \mathbf{m}_j , we need \mathbf{v}_j and the $2p$ previous search directions.

Since \mathbf{R}_j results from the QR-factorization of $\overline{\mathbf{H}}_j$, we can use Givens rotations to annihilate the lower subdiagonal entries of $\overline{\mathbf{H}}_j$. This procedure is a generalization of the one for MINRES; however, since there are p lower subdiagonal entries, each column requires p Givens rotations. At iteration j , let $\mathcal{G}_j^{(j)}$ denote the unitary matrix used to annihilate the lower subdiagonal entries of column j of $\overline{\mathbf{H}}_j$. We write $\mathcal{G}_j^{(j)} = \mathbf{G}_{j+1,j}^{(j)} \mathbf{G}_{j+2,j}^{(j)} \cdots \mathbf{G}_{j+p,j}^{(j)}$ where $\mathbf{G}_{i,j}^{(j)}$ is the matrix applying the Givens rotation to annihilate entry $h_{i,j}$ from $\overline{\mathbf{H}}_j$. The application of the Givens rotations in $\mathcal{G}_j^{(j)}$ will affect columns j to $j+2p$; so, at step j , the rotations in $\mathcal{G}_{j-2p}^{(j)}$ to $\mathcal{G}_{j-1}^{(j)}$ will affect the j th column of $\overline{\mathbf{H}}_j$ after it is constructed, prior to computing the Givens rotations for step j .

We can also use Householder reflections which require about half the work of using Givens rotations. It has been shown that we can effectively store products of a series of Householder reflections as a single matrix, applying them at once, without sacrificing accuracy [10]. Based on the implementation details discussed in Section 4.1 below, using Householder reflections here would be straightforward.

Observe that at every step, we construct a new set of Givens rotations. At step j , $\mathcal{G}_j^{(j)}$ is immediately applied to $\mathcal{G}_{j-1}^{(j)} \cdots \mathcal{G}_1^{(j)} \mathbf{E}_1^{(1)} \mathbf{S}$. Initially $\mathbf{E}_1^{(1)} \mathbf{S}$ is upper triangular. Multiplying by the first set of Givens rotations adds a subdiagonal band. From (3.1), we see that $\mathbf{E}_1^{(j-1)} \mathbf{S}$ is a submatrix of $\mathbf{E}_1^{(j)} \mathbf{S}$. Furthermore, $\mathcal{G}_{j-1}^{(j-1)} \cdots \mathcal{G}_1^{(j-1)} \mathbf{E}_1^{(j-1)} \mathbf{S}$ is contained as the upper block in $\mathcal{G}_{j-1}^{(j)} \cdots \mathcal{G}_1^{(j)} \mathbf{E}_1^{(j)} \mathbf{S}$. Therefore, the first subdiagonal in $\mathcal{G}_1^{(1)} \mathbf{E}_1^{(1)} \mathbf{S}$ exists in $\mathcal{G}_1^{(2)} \mathbf{E}_1^{(2)} \mathbf{S}$. Each further application of Givens rotations adds a new subdiagonal so that $\mathcal{G}_j^{(j)} \cdots \mathcal{G}_1^{(j)} \mathbf{E}_1^{(j)} \mathbf{S}$ has j lower subdiagonal bands. The rotations contained in $\mathcal{G}_j^{(j)}$ only effect rows j to $j+p$ of $\mathcal{G}_{j-1}^{(j)} \cdots \mathcal{G}_1^{(j)} \mathbf{E}_1^{(j)} \mathbf{S}$. Thus we have the relation $\mathbf{T}_j = \begin{bmatrix} \mathbf{T}_{j-1} \\ \mathbf{t}_j^T \end{bmatrix}$ where $\mathbf{t}_j \in \mathbb{C}^p$, and we can update \mathbf{X}_j progressively as an update of \mathbf{X}_{j-1} ,

$$\mathbf{X}_j = \mathbf{X}_{j-1} + \mathbf{m}_j \mathbf{t}_j^T. \tag{4.2}$$

In [19], Saad observes that, as in the single right-hand-side case, a computed residual (also sometimes called the *recursive residual*) is available,

$$\|\mathbf{F}_j^{(i)}\| = \|(\overline{\mathbf{T}}_j^{(i)})_{j+1:j+i}\| \tag{4.3}$$

4.1. Implementation Details. We conclude our description with some notes about practical implementation details. To summarize our storage needs, we must store $2p$ block Lanczos vectors, $2p$ search directions, $2p$ sets of Givens rotations (or Householder reflections), the lower subdiagonal entries of p previous columns, and the j th column of $\overline{\mathbf{H}}_j$ which will be transformed into the j th column of $\overline{\mathbf{R}}_j$. The fact that \mathbf{A} is symmetric allows for large savings in the storage requirements of the block MINRES algorithm, but this also destroys much of the natural indexing that would be available if complete storage were necessary. We want the algorithm to deal with as few special cases as possible. With this in mind, we present data structures which allow for simple, efficient local indexing from which some simple patterns arise, allowing for a clean, simple algorithm with only one special case for the first iteration. We use block size $p = 5$ when presenting examples demonstrating the utility of the strategy. Also, we denote the first-in-first-out (FIFO) queue data structure simply as a *queue*, for convenience.

Storing basis vectors and search direction in queues offers many advantages. In Matlab, this structure does not exist. Thus, we will store information in an array, column-wise and treat it like a queue. A column in the array will represent data generated at a particular iteration. Any initial data is stored in the rightmost columns, and when a new column of information is created, it is inserted into the last column with all other data being shifted to the left by one column. The first column of data is discarded as new data is added to the last column. This means that in any data structure, the last column is associated to the current iteration (the most current information) allowing for local indexing, i.e., indexing relative to the most recently created column.[†]

We must store $2p$ block Lanczos vectors, and the most straightforward structure to store them in is a $n \times 2p$ array called \mathbf{V} . At the beginning, \mathbf{V} is initialized with all zeros, and the columns of \mathbf{U}_p are stored in the last p columns of \mathbf{V} . Aside from providing a natural way to create and discard Lanczos vectors without explicitly keeping track of indices; with this setup, we always know that the vector in the $(p+1)$ st column is the one on which \mathbf{A} acts to generate the next Lanczos vector. The $2p$ search directions are stored in a similar structure called \mathbf{M} , and \mathbf{M} is initialized to all zeros.

For the matrices $\overline{\mathbf{H}}_j$ and \mathbf{R}_j , we only need to store the complete set of nonzero entries of the j th column. The j th column of $\overline{\mathbf{H}}_j$ has at most $2p+1$ nonzero entries. We store them in an array called \mathbf{r} which is large enough to accommodate padding at the top by zeros which will be filled when the Givens rotations from previous steps are applied. While we do not need to store all the entries of the previous columns of $\overline{\mathbf{H}}_j$ or \mathbf{R}_j , we do need to store the lower subdiagonal entries of the p most recent columns of $\overline{\mathbf{H}}_j$, since the block Lanczos method uses the symmetry of \mathbf{H}_j to fill the superdiagonal entries of the newest column. We store these entries in a $p \times p$ array called \mathbf{H} where each column of \mathbf{H} contains the lower subdiagonal entries of a particular column of $\overline{\mathbf{H}}_j$. In the beginning, \mathbf{H} is initialized as all zeros, and the columns of \mathbf{H} are filled starting from the right, functioning as a queue, as we have already described. Storing the entries in this manner results in the nonzero superdiagonal entries of the current

[†]It should be noted that to achieve this queue-like behavior without actually moving columns of data (an expensive memory movement procedure), we store a set of integer indices which act as pointers to the columns of the matrix. These indices are permuted rather than the actual columns. For example, if the i th entry in the index list is k , then column i of the array is actually stored in the k th column of the data structure.

column of $\bar{\mathbf{H}}_j$ being available as the nonzero antidiagonal entries of \mathbf{H} . This allows us to obtain the super diagonal entries of the current column without computing the associated inner products. For example, suppose $j = 7$, so we are constructing the 7th column of $\bar{\mathbf{H}}_j$. Then we would have

$$\mathbf{r} = \begin{bmatrix} 0 \\ \mathbf{h}_{2,7} \\ \mathbf{h}_{3,7} \\ \mathbf{h}_{4,7} \\ \mathbf{h}_{5,7} \\ \mathbf{h}_{6,7} \\ h_{7,7} \\ h_{8,7} \\ h_{9,7} \\ h_{10,7} \\ h_{11,7} \\ h_{12,7} \\ h_{13,7} \\ h_{14,7} \end{bmatrix} \quad \text{and} \quad \mathbf{H} = \begin{bmatrix} h_{3,2} & h_{4,3} & h_{5,4} & h_{6,5} & \mathbf{h}_{7,6} \\ h_{4,2} & h_{5,3} & h_{6,4} & \mathbf{h}_{7,5} & h_{8,6} \\ h_{5,2} & h_{6,3} & \mathbf{h}_{7,4} & h_{8,5} & h_{9,6} \\ h_{6,2} & \mathbf{h}_{7,3} & h_{8,4} & h_{9,5} & h_{10,6} \\ \mathbf{h}_{7,2} & h_{8,3} & h_{9,4} & h_{10,5} & h_{11,6} \end{bmatrix}. \quad (4.4)$$

Observe in (4.4) that the bold entries in \mathbf{r} can be obtained from the antidiagonal entries of \mathbf{H} , also bolded. Though these entries are computed implicitly due to symmetry, one may want to explicitly compute them for increased numerical stability. This is of particular concern in the computation of eigenvalues, but it is still worth mentioning in this context.

Lastly, the Givens sine and cosines are stored in $p \times 2p$ arrays \mathbf{s} and \mathbf{c} . Each column of \mathbf{s} and \mathbf{c} represents the sines and cosines of Givens rotations used to annihilate entries of a particular column $\bar{\mathbf{H}}_j$. We present Algorithm 4.1, which describes the Ruhe version of block MINRES, but without the implementation tricks just described.

4.1.1. Removal of Dependent Basis Vectors. In the single-vector GMRES minimum residual method (and thus the MINRES method), it may happen that at step j , we have that $\mathbf{A}\mathbf{v}_j \in \mathbb{K}_j(\mathbf{A}, \mathbf{r}_0)$. This situation is referred to as *happy breakdown* since it means that the true solution is contained in the existing Krylov subspace, see, e.g., [19, Section 6.5.4]. Unfortunately, this is not the case in a block Krylov subspace method. We may have dependent block Krylov subspace basis vectors without convergence any of the systems. In the case of algorithms which are built upon the symmetric or nonsymmetric block Lanczos methods, this dependence of the basis can lead to unstable algorithms if not properly handled; see, e.g., [7, 13].

Various strategies have been proposed to mitigate the basis dependence problem. For block-level algorithms, one must first compute or estimate the range of the block Krylov subspace basis to detect rank deficiency. For symmetric Lanczos-based methods, O’Leary [13] advocates removal of the dependent vector, reducing the block size. The update procedures for the systems not associated to the removed vector do not change, and a progressive update formula can be derived for the systems associated to removed right-hand sides. Baglama [3] suggests that instead of simply removing the dependent vector and reducing block size, one can instead replace the dependent vector with a random vector which has been orthogonalized against all previous Lanczos vectors and continue unabated. For nonsymmetric Lanczos-based block QMR, Aliaga et al. [2] propose to remove basis vectors before dependence is detected. Due to issues of stability in block nonsymmetric Lanczos based methods, the authors advocate

defining a tolerance $d_{tol} > 0$. After a vector \mathbf{v} has been biorthogonalized we have

Algorithm 4.1: Block MINRES (Ruhe Implementation)

Input : $\mathbf{A} \in \mathbb{C}^{n \times n}$ Symmetric, $\mathbf{B} \in \mathbb{C}^{n \times p}$, $\mathbf{X}_0 = \mathbf{0}$, $\epsilon > 0$, $M \in \mathbb{N}$

Output: $\mathbf{X} \in \mathbb{C}^{n \times p}$ such that

$$\|\mathbf{B}(:, j) - \mathbf{A}\mathbf{X}(:, j)\| / \|\mathbf{B}(:, j) - \mathbf{A}\mathbf{X}_0(:, j)\| < \epsilon \quad \forall j \leq p$$

1 Compute the QR-Factorization $\mathbf{B}\mathbf{V}_p\mathbf{S}$

2 $\hat{\mathbf{S}} \leftarrow \mathbf{S}\mathbf{E}_1^{(1)}$

3 $\mathbf{X} \leftarrow \mathbf{X}_0$

4 $\mathbf{R} \leftarrow \mathbf{B} - \mathbf{A}\mathbf{X}$

5 **while** $\max_{0 < i \leq p} \left\{ \left\| (\overline{\mathbf{T}}_j^{(i)})_{j+1:j+i} \right\| \right\} < \epsilon \|\mathbf{b}^{(i)}\|$ **and** $j \leq M$ **do**

6 $\mathbf{w} \leftarrow \mathbf{A}\mathbf{v}_j$

7 **if** $j > 1$ **then**

8 **for** $i = j - p : j - 1$ **do**

9 $h_{i,j} = h_{j,i}$

10 $\mathbf{w} \leftarrow h_{i,j}\mathbf{w}$

11 **for** $i = j : j + p - 1$ **do**

12 $h_{i,j} = \mathbf{v}_i^* \mathbf{w}$

13 $\mathbf{w} \leftarrow h_{i,j}\mathbf{v}_i$

14 $h_{j+p,j} = \|\mathbf{w}\|$

15 $\mathbf{v}_{j+1} = \mathbf{w}/h_{j+p,j}$

16 **if** $j > 1$ **then**

17 $\overline{\mathbf{r}}_j^{(j)} \leftarrow \mathcal{G}_{j-1}^{(j)} \cdots \mathcal{G}_{j-2p}^{(j)} \overline{\mathbf{h}}_j^{(j)}$

18 $\mathcal{G}_j^{(j)} \leftarrow \mathbf{I}_j$

19 **for** $i = j + p - 1 : j$ **do**

20 Generate Givens rotation for $\mathbf{G}_{i+1,j}^{(j)}$ to annihilate $h_{i+1,j}$

21 $\overline{\mathbf{r}}_j^{(j)} \leftarrow \mathbf{G}_{i+1,j}^{(j)} \overline{\mathbf{r}}_j^{(j)}$

22 $\hat{\mathbf{S}} \leftarrow \mathbf{G}_{i+1,j}^{(j)} \hat{\mathbf{S}}$

23 $\mathcal{G}_j^{(j)} \leftarrow \mathbf{G}_{i+1,j}^{(j)} \mathcal{G}_j^{(j)}$

24 **if** $m = 1$ **then**

25 $\mathbf{m}_1 = \mathbf{v}_1 / \overline{\mathbf{R}}_1(1, 1)$

26 **else**

27 $\mathbf{w} \leftarrow \mathbf{v}_j$

28 **for** $i = j - 2p : j - 1$ **do**

29 $\mathbf{w} \leftarrow \mathbf{w} - \overline{\mathbf{R}}_j(i, j)\mathbf{m}_i$

30 $\mathbf{m}_j = \mathbf{w} / \overline{\mathbf{R}}_j(j, j)$

31 $\mathbf{t}^T \leftarrow \hat{\mathbf{S}}(j, :)$

32 $\mathbf{X} \leftarrow \mathbf{X} + \mathbf{m}_j \mathbf{t}^T$

33 $\hat{\mathbf{S}} \leftarrow \begin{bmatrix} \hat{\mathbf{S}} \\ \mathbf{0}_{1 \times p} \end{bmatrix}$

34 $j \leftarrow j + 1$

$\|\mathbf{v}\| < d_{tol}$, then we consider \mathbf{v} as almost being dependent, and it is removed from the basis. In [2], a bookkeeping scheme is presented to keep track of such removals so that the block QMR algorithm can be adjusted accordingly. Recently, this technique was extended to a block conjugate gradient method for shifted linear systems [4]. The book keeping scheme allows for the dependent basis vectors to be removed from the block Lanczos process but temporarily retained in memory for the purposes of orthogonalization.

Dubrulle [6] proposes an alternative to the removal of dependent or near-dependent basis vectors, for use in a block conjugate gradient algorithm. He proposes to use a change-of-basis strategy for the block descent directions and other algorithmic changes to avoid the problem long before near-rank deficiency of the block basis vectors occurs. This additionally avoids the need for basis rank estimation.

For our version of the block MINRES algorithm, we act only when an actual dependency occurs (where dependency has the numerical definition that $h_{j+p,j} < \gamma$ for a pre chosen $0 < \gamma \ll 1$). One of the strengths of a block MINRES algorithm built from Ruhe's block Lanczos method is that there is no need for any basis rank estimation. Since we construct only one block Lanczos vector at a time, we simply need to compute $h_{j+p,j}$, i.e., compute the norm of the newest basis vector after orthogonalization via Ruhe's block Lanczos process. This observation has been previously made (in the context of eigenvalue computations) [3]. Baglama presents two options for dealing with basis dependence. One option is to reduce block size by one and adjust short-term recurrences accordingly. The other is to generate a random vector and orthogonalize it with respect to all previous Lanczos vectors. This normalized vector is then put in the place of the dependent basis vector. We show that either option results in minimal changes to either Algorithm 4.1 or the implementation details presented in the previous section. We discuss the algorithmic modifications required to incorporate both options into our block MINRES algorithm, but we choose the latter due the simplicity of incorporating it into the algorithm. For illustration, we present examples with smaller block size than previously, for ease of discussion; however, it is clear that the simplifications presented do not change if the block size increases.

We discuss, first, the effects of choosing the first option, block size reduction. Suppose we have block size $p = 2$, and we run ten iterations of the Ruhe block Lanczos algorithm, such that no dependent basis vectors arise. We have generated block Krylov subspace $\mathbb{K}_{12,6,0}(\mathbf{A}, \mathbf{U}_2)$. Implicitly, we have also generated

$$\bar{\mathbf{H}}_{10} = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & & & & & & & & & \\ h_{2,1} & h_{2,2} & h_{2,3} & h_{2,4} & & & & & & & & \\ h_{3,1} & h_{3,2} & h_{3,3} & h_{3,4} & h_{3,5} & & & & & & & \\ & h_{4,2} & h_{4,3} & h_{4,4} & h_{4,5} & h_{4,6} & & & & & & \\ & & h_{5,3} & h_{5,4} & h_{5,5} & h_{5,6} & h_{5,7} & & & & & \\ & & & h_{6,4} & h_{6,5} & h_{6,6} & h_{6,7} & h_{6,8} & & & & \\ & & & & h_{7,5} & h_{7,6} & h_{7,7} & h_{7,8} & h_{7,9} & & & \\ & & & & & h_{8,6} & h_{8,7} & h_{8,8} & h_{8,9} & h_{8,10} & & \\ & & & & & & h_{9,7} & h_{9,8} & h_{9,9} & h_{9,10} & & \\ & & & & & & & h_{10,8} & h_{10,9} & h_{10,10} & & \\ & & & & & & & & h_{11,9} & h_{11,10} & & \\ & & & & & & & & & h_{12,10} & & \end{bmatrix} \quad (4.5)$$

where $\bar{\mathbf{H}}_{10}$ is symmetric, and we have $\mathbf{A}\mathbf{U}_{10} = \mathbf{U}_{12}\bar{\mathbf{H}}_{10}$. Now, suppose that instead

$\mathbf{A}\mathbf{u}_5 \in \mathbb{K}_{6,3,0}(\mathbf{A}, \mathbf{U}_2)$, numerically. This means that

$$\mathbf{A}\mathbf{u}_5 = h_{7,5}\mathbf{u}_7 + h_{6,5}\mathbf{u}_6 + h_{5,5}\mathbf{u}_5 + h_{4,5}\mathbf{u}_4 + h_{3,5}\mathbf{u}_3 \quad (4.6)$$

with $h_{7,5} < \gamma$. Then we take $\mathbf{u}_7 = \mathbf{0}$, do not add it to our basis, and assign $h_{7,5} = 0$. Furthermore, since this implies that $\mathbf{A}\mathbf{u}_7 = \mathbf{0}$, we have that $\mathbf{u}_9 = \mathbf{0}$, $\mathbf{u}_{11} = \mathbf{0}$, etc. Thus, certain entries of $\overline{\mathbf{H}}_{10}$ will be annihilated. Specifically, columns seven and nine are now zero. By symmetry, nonzero entries in rows seven and nine are also annihilated, and the same is true for row eleven. With this in mind, we can actually construct a smaller, banded matrix and write a more compact block Lanczos relation which ignores the eliminated Lanczos vectors. Note that for the purposes of this description, we do not renumber the block Lanczos vectors when one is eliminated due to the removal of a dependent basis vector. We maintain the same indexing as before, meaning indices associated to annihilated vectors are no longer used. Let

$$\tilde{\mathbf{H}}_{10} = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & & & & & & & \\ h_{2,1} & h_{2,2} & h_{2,3} & h_{2,4} & & & & & & \\ h_{3,1} & h_{3,2} & h_{3,3} & h_{3,4} & h_{3,5} & & & & & \\ & h_{4,2} & h_{4,3} & h_{4,4} & h_{4,5} & h_{4,6} & & & & \\ & & h_{5,3} & h_{5,4} & h_{5,5} & h_{5,6} & & & & \\ & & & h_{6,4} & h_{6,5} & h_{6,6} & h_{6,8} & & & \\ & & & & & h_{8,6} & h_{8,8} & h_{8,10} & & \\ & & & & & & h_{10,8} & h_{10,10} & & \\ & & & & & & & h_{12,10} & & \end{bmatrix} \quad (4.7)$$

so that we have the compacted block Lanczos relation $\mathbf{A}\tilde{\mathbf{U}}_{10} = \tilde{\mathbf{U}}_{12}\tilde{\mathbf{H}}_{10}$, where

$$\tilde{\mathbf{U}}_{12} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_6 \quad \mathbf{u}_8 \quad \mathbf{u}_{10} \quad \mathbf{u}_{12}],$$

and $\tilde{\mathbf{U}}_{10}$ is similarly defined but without the last column. Notice that this removal of one dependent basis vector results in a reduction of the effective bandwidth by two when we switch from $\overline{\mathbf{H}}_{10}$ to $\tilde{\mathbf{H}}_{10}$. However, this reduction happens in two stages. First, bandwidth reduces once in column five. Then in column eight (actually the seventh column) there is a second reduction. This can easily be generalized. For block size p , if we observe that $\mathbf{A}\mathbf{u}_i$ is contained in an old Krylov subspace and therefore $\mathbf{u}_{i+p} = \mathbf{0}$, then we will see a reduction of bandwidth at column i and another reduction at column $i+p$. This implies a reduction in the number of vectors required for storage in the block Lanczos process. We need two fewer vectors for each block size reduction.

This change in bandwidth is reflected in the upper triangular bands of $\tilde{\mathbf{R}}_{10}$, the upper triangular matrix defined by the QR factorization $\tilde{\mathbf{H}}_{10} = \tilde{\mathbf{Q}}_{10}\tilde{\mathbf{R}}_{10}$. We have the structure

$$\tilde{\mathbf{R}}_{10} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & r_{1,5} & & & & & \\ & r_{2,2} & r_{2,3} & r_{2,4} & r_{2,5} & \mathbf{0} & & & & \\ & & r_{3,3} & r_{3,4} & r_{3,5} & r_{3,6} & & & & \\ & & & r_{4,4} & r_{4,5} & r_{4,6} & r_{4,8} & & & \\ & & & & r_{5,5} & r_{5,6} & r_{5,8} & \mathbf{0} & & \\ & & & & & r_{6,6} & r_{6,8} & r_{7,10} & & \\ & & & & & & r_{8,8} & r_{8,10} & & \\ & & & & & & & r_{10,10} & & \end{bmatrix}, \quad (4.8)$$

with bolded zeros indicating where an entry was annihilated by the reduction in block size. Recalling (4.1), this indicates that our storage requirements for constructing the search directions will change as the bandwidth changes. To construct \mathbf{m}_5 , we need to store \mathbf{m}_4 , \mathbf{m}_3 , \mathbf{m}_2 , and \mathbf{m}_1 along with \mathbf{u}_5 . However, to construct \mathbf{m}_6 , we only need to store four vectors: \mathbf{m}_5 , \mathbf{m}_4 , \mathbf{m}_3 , and \mathbf{u}_6 . For \mathbf{m}_{10} , we only require three vectors: \mathbf{m}_8 , \mathbf{m}_7 , and \mathbf{u}_{10} . This can easily be generalized for block- p . For each dependent basis vector removed, we will have a two-vector reduction in storage requirements for the construction of the search directions. In total, for each block size reduction, we have a four-vector reduction in storage requirements.

We now discuss how choosing the second option (inserting a random, orthogonalized vector into the basis) affects the algorithm. We begin by describing the replacement procedure in more detail. At iteration j , we compute $\mathbf{A}\mathbf{u}_j$. After orthogonalization, we see that $h_{j+p,j} < \gamma$. Thus, $\mathbf{A}\mathbf{u}_j$ is (numerically) in the range of the previous Lanczos vectors. We set $h_{j+p,j} = 0$ thus the coefficients in column j encode that $\mathbf{A}\mathbf{u}_j \in \text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_j\}$. Let $\hat{\mathbf{u}}_{j+p}$ be a vector constructed by taking a random vector $\hat{\mathbf{w}}$, orthogonalizing $\hat{\mathbf{w}}$ with respect to all previous Lanczos vectors, and setting $\hat{\mathbf{u}}_{j+p} = \hat{\mathbf{w}}/\|\hat{\mathbf{w}}\|$. Then the algorithm continues as before with this modified block Lanczos basis.

What modifications must be made to the block MINRES algorithm to accommodate this strategy? It turns out, very few. Of course, we do not store the complete Lanczos basis, as this would defeat the purpose of developing a method for symmetric systems. However, we need to orthogonalize the random vector against the entire basis. As a work-around, we can generate a random vector at the start of the iteration and simply orthogonalize against each Lanczos vector as it is created. This would require only one additional vector of storage and an additional orthogonalization per iteration. If we are solving a problem in which we expect there to be more than one occurrence of loss of linear independence, we can generate more than one random vector, balancing between increasing the storage requirements and insuring against the basis dependence problem.

One might be concerned that introducing a vector not created by the block Lanczos process will destroy the short-term recurrences which make symmetric Lanczos methods so attractive. However, this is not the case. Suppose that after iteration j , we continue Ruhe's block Lanczos process with the modified basis. Let $\hat{\mathbf{U}}_{j+p} \in \mathbb{C}^{n \times (j+p)}$ be the matrix containing the block Lanczos vectors but with $\hat{\mathbf{u}}_{j+p}$ as its last column. Observe that the matrix $\hat{\mathbf{U}}_{j+p}^* \mathbf{A} \hat{\mathbf{U}}_{j+p}$ is symmetric; inserting the new basis vector does not effect this. Thus, the banded structure of $\hat{\mathbf{H}}_j$ defined by $\mathbf{A} \hat{\mathbf{U}}_j = \hat{\mathbf{U}}_{j+p} \hat{\mathbf{H}}_j$ is the same as that of $\bar{\mathbf{H}}_j$. The only change is that we now have zero entries at $h_{j+p,j}$ and $h_{j,j+p}$. This, in turn, gives a slight change in structure to $\hat{\mathbf{R}}_j$, the upper triangular factor in the QR-factorization of $\hat{\mathbf{H}}_j$.

As an example, suppose $p = 2$ and that $\mathbf{A}\mathbf{v}_5$ is in the span of the existing block Lanczos vectors, as in the last example. If we continue the Ruhe's block Lanczos process with the modified basis, we have the following structures for $\hat{\mathbf{H}}_8$, and $\hat{\mathbf{R}}_8 \in \mathbb{C}^{10 \times 8}$, the upper triangular factor in the QR-factorization of $\hat{\mathbf{H}}_8$ constructed using

Givens rotations,

$$\hat{\mathbf{H}}_8 = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & & & & & & & \\ h_{2,1} & h_{2,2} & h_{2,3} & h_{2,4} & & & & & & \\ h_{3,1} & h_{3,2} & h_{3,3} & h_{3,4} & h_{3,5} & & & & & \\ & h_{4,2} & h_{4,3} & h_{4,4} & h_{4,5} & h_{4,6} & & & & \\ & & h_{5,3} & h_{5,4} & h_{5,5} & h_{5,6} & \mathbf{0} & & & \\ & & & h_{6,4} & h_{6,5} & h_{6,6} & h_{6,7} & h_{6,8} & & \\ & & & & \mathbf{0} & h_{7,6} & h_{7,7} & h_{7,8} & h_{7,9} & \\ & & & & & h_{8,6} & h_{8,7} & h_{8,8} & h_{8,9} & h_{8,10} \\ & & & & & & h_{9,7} & h_{9,8} & h_{9,9} & h_{9,10} \\ & & & & & & & h_{10,8} & h_{10,9} & h_{10,10} \\ & & & & & & & & h_{11,9} & h_{11,10} \\ & & & & & & & & & h_{12,10} \end{bmatrix}.$$

and

$$\widehat{\mathbf{R}}_8 = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & r_{1,5} & & & & & & \\ & r_{2,2} & r_{2,3} & r_{2,4} & r_{2,5} & r_{2,6} & & & & & \\ & & r_{3,3} & r_{3,4} & r_{3,5} & r_{3,6} & \mathbf{0} & & & & \\ & & & r_{4,4} & r_{4,5} & r_{4,6} & r_{4,7} & r_{4,8} & & & \\ & & & & r_{5,5} & r_{5,6} & r_{5,7} & r_{5,8} & \mathbf{0} & & \\ & & & & & r_{6,6} & r_{6,7} & r_{6,8} & r_{6,9} & r_{6,10} & \\ & & & & & & r_{7,7} & r_{7,8} & r_{7,9} & r_{7,10} & \\ & & & & & & & r_{8,8} & r_{8,9} & r_{8,10} & \\ & & & & & & & & r_{9,9} & r_{9,10} & \\ & & & & & & & & & r_{10,10} & \end{bmatrix}$$

This indicates that the final effects of replacing the dependent basis vector with a random one are minimal. The two zeros are introduced into upper Hessenberg matrix, but the bandwidth and symmetry properties remain unchanged. The introduction of a zero in the seventh column of $\mathbf{\tilde{R}}_8$ and another in the ninth simply means that the seventh and ninth block Lanczos vectors are linear combinations of the previous four rather than the previous five search directions, recalling the construction of the search directions (4.1).

5. Convergence Theory. Theoretically, MINRES is a version of block GMRES for symmetric systems. Simoncini and Gallopoulos discussed the convergence properties of block GMRES [21], including a result by Vital [23]. We can easily describe the quality of the residual produced at iteration $j = k + mp$. For $\mathbf{b}^{(i)}$, the i th column of the right-hand side \mathbf{B} , Algorithm 4.1 minimizes the i th column of the residual $\mathbf{f}_j^{(i)}$ over the subspace $\mathbb{K}_{j,k,m}(A, \mathbf{F}_0)$. Thus, we can expect $\|\mathbf{f}_j^{(i)}\|$ to be at least as good as the norm of the residual produced by running J steps of MINRES with $\mathbf{b}^{(i)}$ as the single right-hand side, where $J = \begin{cases} k & \text{if } i < m \\ k + 1 & \text{if } i \geq m \end{cases}$. This easily can be understood by recalling the definition of $\mathbb{K}_{j,k,m}(\mathbf{A}, \mathbf{F}_0)$ in (2.3). We observe that having a larger subspace over which to minimize is not guaranteed to give dramatic improvements in convergence. The additional information contained in $\mathbb{K}_{j,k,m}(A, \mathbf{F}_0)$ may not be

helpful in the minimization process. For specially related right-hand sides, though, we may have convergence in many fewer iterations.

6. Numerical Results. These numerical experiments are meant to demonstrate the effectiveness and behavior of Algorithm 4.1. In all experiments, we compare the performance of block MINRES with sequential applications of Matlab’s MINRES function. We compared performance using iteration counts and CPU timings. All tests were performed on a Macbook Pro containing a 2.3 GHz Intel Core i5 processor with 8 GB of 1333MHz DDR3 main memory running the 64-bit version of Matlab R2011b. In any experiment involving the generation of random vectors, we used Matlab’s `mt19937ar` random number generator, with seed 0, which was initialized at the beginning of each experiment. Let $\mathbf{L} \in \mathbb{C}^{n_1 \times n_1}$, with $n_1 = 40000$, be the discretization of the Laplacian operator on a 200×200 regular grid using central differences. This matrix is negative-definite. Let $\mathbf{A} = -\mathbf{L} + 200\mathbf{I}$. Due to the eigenvalue distribution of \mathbf{L} , we have that \mathbf{A} is indefinite.

It should be noted that using Ruhe’s Lanczos implementation allows us to compare with sequential applications of a single-vector method iteration for iteration. This is particularly useful if we are using a block method simply to accelerate the convergence of an iteration applied to a system with one right-hand side, i.e., we have generated a few random artificial right-hand sides to create the block Krylov subspace.

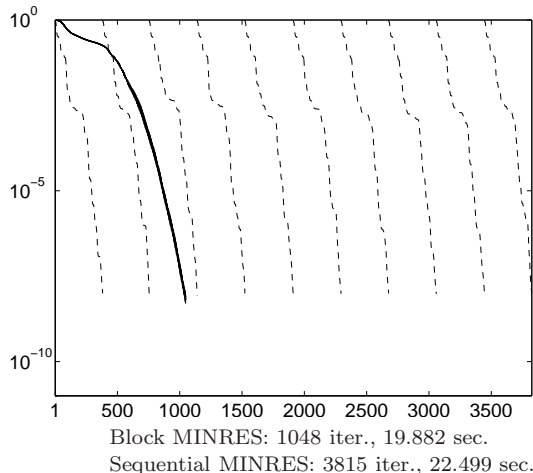


FIG. 6.1. Comparison of the performance of Algorithm 4.1 versus sequential applications of MINRES on the discretized Laplacian system with ten randomly generated right-hand sides. The **solid black** curve is actually the ten convergence curves for each right-hand side when solved by Algorithm 4.1 overlaid on one another. We see that in the case of these ten right-hand sides that block MINRES convergence for all ten systems is qualitatively the same. The **black dashed** curves are the convergence curves for each sequential application of MINRES for each right-hand side.

We begin by demonstrating the performance of the algorithm on the shifted Laplacian system with ten randomly generate right-hand sides. In Figure 6.1, we see that for these right-hand sides, the block MINRES algorithm offers a performance improvement in terms of iterations and in time. The time improvement is moderate, but for larger, more-expensive-to-apply operators, the improvement in time could be more substantial.

We demonstrate that our removal of dependent basis vectors works as described. Of course, it is difficult to choose a pair of right-hand sides for which basis dependence

will occur in later iterations. Thus, as a simple, easy-to-construct test, we choose the first right-hand side \mathbf{e}_1 , as the first canonical basis vector. The second right-hand side is $\mathbf{A}\mathbf{e}_1$, the image of the first canonical basis vector, i.e., the first column of our coefficient matrix. This will result in basis dependence at the first iteration of our algorithm. As is shown in Figure 6.2, this leads to immediate convergence for that system when running block MINRES. Of course, this example is not likely to occur in practice. It merely demonstrates that the algorithm can handle dependence gracefully.

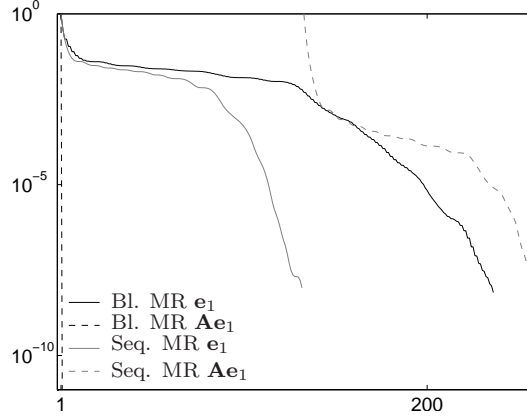
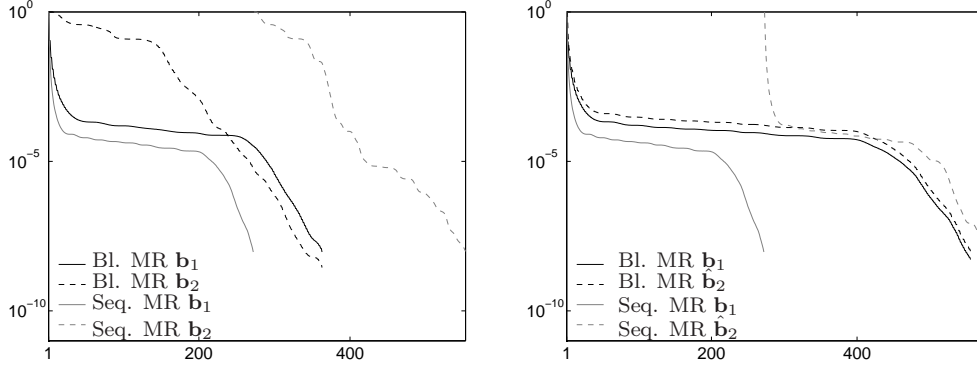


FIG. 6.2. *Demonstration of the algorithm's performance in the case that it encounters basis dependence. In this case, with the right-hand sides \mathbf{e}_1 and $\mathbf{A}\mathbf{e}_1$, dependence occurs at the first iteration. Since the first right-hand side is the solution to the second system, we get immediate convergence for the second system, and block MINRES continues for the other system, replacing the dependent basis vector with a random one.*

We demonstrate how the relationship between the right-hand sides can effect the performance of this implementation of block MINRES with three experiments.

We compared the performance of our block MINRES implementation with that of sequential runs of Matlab's MINRES for \mathbf{A} with three pairs of right-hand sides. For the first pair, let $\mathbf{b}_1 = \mathbf{e}_{n_1}^{(1)}$ and $\mathbf{b}_2 = \mathbf{1}$, the vector of all ones. For second pair of right-hand sides, we let $\hat{\mathbf{b}}_1 = \mathbf{b}_1$ but change the second right-hand side by letting $\hat{\mathbf{b}}_2 = \mathbf{e}_{n_1}^{(2)}$. In Figure 6.3, we show a comparison of convergence curves for these pairs of right-hand sides. We observe that exchanging \mathbf{b}_2 for $\hat{\mathbf{b}}_2$ degrades the performance of our Block MINRES implementation. Recall that the convergence of a Krylov subspace method for a Hermitian system is completely determined by its eigenvalues. For an indefinite system, the eigenvalues closest to the origin cause a delay in convergence. Therefore, we hypothesize that a pair of right-hand sides that have strong components from different parts of the eigenspace associated to eigenvalues of small magnitude might complement each other well. Let $\mathbf{Q} \in \mathbb{C}^{n_1 \times 50}$ have orthonormal columns spanning the eigenspace associated to the 50 eigenvalues of \mathbf{A} of smallest magnitude. We can study the magnitude of the components of \mathbf{b}_1 , \mathbf{b}_2 , and $\hat{\mathbf{b}}_2$ in $\mathcal{R}(\mathbf{Q})$ to better understand the convergence curves we see in Figure 6.3. In Figure 6.4, we plot the components of $\mathbf{Q}^*\mathbf{b}_1$, $\mathbf{Q}^*\mathbf{b}_2$, and $\mathbf{Q}^*\hat{\mathbf{b}}_2$. We see that the eigencomponents of \mathbf{b}_1 , and $\hat{\mathbf{b}}_2$ are similar in magnitude, meaning that a block Krylov subspace for these two right-hand sides might not contain subspace information useful for accelerating convergence, when



Block MINRES: 362 iter., 2.2867 sec.
 Sequential MINRES: 551 iter., 3.2623 sec.

Block MINRES: 555 iter., 3.4629 sec.
 Sequential MINRES: 572 iter., 3.3927 sec.

FIG. 6.3. Performance of block MINRES for different right-hand sides. In the figure on the left, the two right-hand sides are $\mathbf{b}_1 = \mathbf{e}_{n_1}^{(1)}$ and $\mathbf{b}_2 = \mathbf{1}$. In the figure on the right, \mathbf{b}_1 does not change, but $\hat{\mathbf{b}}_2 = \mathbf{e}_{n_1}^{(2)}$.

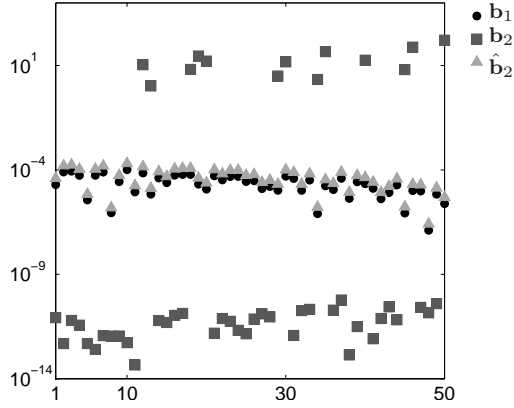


FIG. 6.4. Magnitude of the components of different right-hand-sides in the eigenspace associated to the fifty eigenvectors associated to the smallest magnitude eigenvalues.

compared to its single-vector Krylov subspace counterparts. Many of the components of \mathbf{b}_2 in $\mathcal{R}(\mathbf{Q})$ are orders of magnitude smaller than those of \mathbf{b}_1 . Thirteen are orders of magnitude larger. This may be part of the reason that we are able to achieve some acceleration of convergence using block MINRES for this pair of right-hand-sides. To test this theory, we can construct a new right-hand-side $\hat{\mathbf{b}}_2 = \mathbf{b}_2 - \mathbf{Q}\mathbf{Q}^*\mathbf{b}_2 + \mathbf{Q}\mathbf{Q}^*\hat{\mathbf{b}}_2$. This has the effect of removing the components of \mathbf{b}_2 in $\mathcal{R}(\mathbf{Q})$ and replacing them with the components of $\hat{\mathbf{b}}_2$. In Figure 6.5, we see that when applying block MINRES to the pair of right-hand-sides \mathbf{b}_1 and $\hat{\mathbf{b}}_2$, the method is less effective when compared to the performance \mathbf{b}_1 and \mathbf{b}_2 in Figure 6.3.

This is by no means a rigorous analysis of convergence of a block method. These experiments only are meant to illustrate the variability of performance of a block

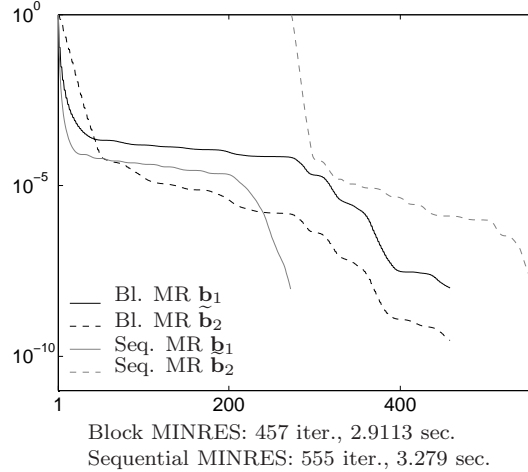


FIG. 6.5. Performance of block MINRES after altering some components of the right-hand side \mathbf{b}_2 .

method for different right-hand sides and provide some insight into this phenomenon.

7. Conclusions. We have presented an alternative implementation of the block MINRES residual algorithm. This version is based on Ruhe's block Krylov subspace basis generation strategy which produces one orthonormal basis vector at a time rather than a full block of vectors. We provide not only a theoretical derivation of the algorithm but also a discussion of the practical implementation issues which need to be addressed to fully take advantage of the efficiencies which arise in a block method for symmetric systems. This variant of the block MINRES method handles dependence of block Krylov subspace basis vectors in a more graceful manner than its block-level brethren. A software implementation in Matlab is provided at <http://math.soodhalter.com/software.php>.

Acknowledgement. The author would like to thank Sebastian Birk, Michael Parks, and Daniel Szyld for their constructive editorial comments and suggestions.

REFERENCES

- [1] A. M. ABDEL-REHIM, R. B. MORGAN, D. A. NICELY, AND W. WILCOX, *Deflated and restarted symmetric Lanczos methods for eigenvalues and linear equations with multiple right-hand sides*, SIAM J. Sci. Comput., 32 (2010), pp. 129–149.
- [2] J. I. ALIAGA, D. L. BOLEY, R. W. FREUND, AND V. HERNÁNDEZ, *A Lanczos-type method for multiple starting vectors*, Mathematics of computation, 69 (2000), pp. 1577–1602.
- [3] J. BAGLAMA, *Dealing with linear dependence during the iterations of the restarted block Lanczos methods*, Numer. Algorithms, 25 (2000), pp. 23–36.
- [4] S. BIRK AND A. FROMMER, *A deflated conjugate gradient method for multiple right-hand sides and multiple shifts*, (In preparation).
- [5] T. F. CHAN AND W. L. WAN, *Analysis of projection methods for solving linear systems with multiple right-hand sides*, SIAM Journal on Scientific Computing, 18 (1997), pp. 1698–1721.
- [6] A. A. DUBRULLE, *Retooling the method of block conjugate gradients*, Electron. Trans. Numer. Anal., 12 (2001), pp. 216–233 (electronic).
- [7] R. W. FREUND AND N. M. NACHTIGAL, *QMR: a quasi-minimal residual method for non-Hermitian linear systems*, Numerische Mathematik, 60 (1991), pp. 315–339.
- [8] A. GREENBAUM, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

- [9] M. H. GUTKNECHT, *Block Krylov space methods for linear systems with multiple right-hand sides: an introduction*, in Modern Mathematical Models, Methods and Algorithms for Real World Systems, A. H. Siddiqi, I. S. Duff, and O. Christensen, eds., New Delhi, 2007, Anamaya Publishers, pp. 420–447.
- [10] M. H. GUTKNECHT AND T. SCHMELZER, *Updating the QR decomposition of block tridiagonal and block Hessenberg matrices*, Applied Numerical Mathematics, 58 (2008), pp. 871–883.
- [11] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, Journal of Research of the National Bureau of Standards, 49 (1952), pp. 409–436 (1953).
- [12] A. A. NIKISHIN AND A. Y. YEREMIN, *Variable block CG algorithms for solving large sparse symmetric positive definite linear systems on parallel computers, I: general iterative scheme*, SIAM Journal on Matrix Analysis and Applications, 16 (1995), p. 1135.
- [13] D. P. O’LEARY, *The block conjugate gradient algorithm and related methods*, Linear Algebra and its Applications, 29 (1980), pp. 293–322.
- [14] C. C. PAIGE AND M. A. SAUNDERS, *Solutions of sparse indefinite systems of linear equations*, SIAM Journal on Numerical Analysis, 12 (1975), pp. 617–629.
- [15] M. L. PARKS, E. DE STURLER, G. MACKEY, D. D. JOHNSON, AND S. MAITI, *Recycling Krylov subspaces for sequences of linear systems*, SIAM Journal on Scientific Computing, 28 (2006), pp. 1651–1674.
- [16] M. L. PARKS, R. SAMPATH, AND P. K. V. V. NUKALA, *Efficient simulation of large-scale 3d fracture networks via krylov subspace recycling*, (In Preparation).
- [17] A. RUHE, *Implementation aspects of band Lanczos algorithms for computation of eigenvalues of large sparse symmetric matrices*, Mathematics of Computation, 33 (1979), pp. 680–687.
- [18] Y. SAAD, *On the Lanczos method for solving symmetric linear systems with several right-hand sides*, Mathematics of Computation, 48 (1987), pp. 651–662.
- [19] ———, *Iterative methods for sparse linear systems*, SIAM, Philadelphia, Second ed., 2003.
- [20] Y. SAAD, M. YEUNG, J. ERHEL, AND F. GUYOMARCH, *A deflated version of the conjugate gradient algorithm*, SIAM Journal on Scientific Computing, 21 (2000), pp. 1909–1926. Iterative methods for solving systems of algebraic equations (Copper Mountain, CO, 1998).
- [21] V. SIMONCINI AND E. GALLOPOULOS, *Convergence properties of block GMRES and matrix polynomials*, Linear Algebra and its Applications, 247 (1996), pp. 97–119.
- [22] C. F. SMITH, A. F. PETERSON, AND R. MITTRA, *A conjugate gradient algorithm for treatment of multiple incident electromagnetic fields*, IEEE Transactions on Antennas and Propagation, 37 (1989), pp. 1490–1493.
- [23] B. VITAL, *Etude de quelques méthodes de résolution de problèmes linéaires de grande taille sur multiprocesseur*, PhD thesis, Université de Rennes, 1990.
- [24] S. WANG, E. DE STURLER, AND G. H. PAULINO, *Large-scale topology optimization using preconditioned Krylov subspace methods with recycling*, International Journal for Numerical Methods in Engineering, 69 (2007), pp. 2441–2468.